ITC 3/53

Information Technology and Control
Vol. 53 / No. 3/ 2024
pp. 695-709
DOI 10.5755/j01.itc.53.3.36121

Six-Degree-of-Freedom Pose Estimation of Class-Level Objects Based on P2T-Net

Received 2024/01/19 Accepted after revision 2024/05/14

HOW TO CITE: Li, G. (2024). Six-Degree-of-Freedom Pose Estimation of Class-Level Objects Based on P2T-Net. *Information Technology and Control*, 53(3), 695-709. https://doi.org/10.5755/j01.itc.53.3.36121

Six-Degree-of-Freedom Pose Estimation of Class-Level Objects Based on P2T-Net

Guanjin Li

Revelle College, University of California San Diego, San Diego, CA 92092, US; e-mail:gul007@ucsd.edu

Corresponding author: gul007@ucsd.edu

6-DoF Pose Estimation (Six-Degree-of-Freedom Pose Estimation) of objects is widely used in the fields of augmented reality, robot operation, and unmanned driving. Due to the complexity and variability of real application scenarios, its task needs to deal with the interference such as light change, distance change, sensor noise, and mutual occlusion of chaotic placement. In application scenarios, the implementation of methods with low hardware cost and also high efficiency on accuracy and time cost is still a challenging problem. At this time, it is important to recognize the class of the object, determine the area of the object in the image, and 6-DoF pose estimation of the object that are still challenging problems. In this paper, we proposed a conceptually simple and data-efficient category-level 6-DoF pose estimation network using Pyramid Pooling Transformer as the foundation network to enhance the accuracy in image classification, semantic segmentation, object detection, and instance segmentation with low hardware cost application background. In the cross-modal fusion phase, the implicit Deep recovery technique is used to improve the RGB-D feature representation capability, and the compact pyramid refinement operation can efficiently fuse multiple layers of features with high speed and few parameters. Compared with traditional methods, the methods we proposed have better resistance to occlusion, MAP of 10° 2cm and 10° 5cm can reach 81.4% and 87.1%, and MAP of 5° 2cm and 5° 5cm can reach 69.2% and 72.9%, which is ahead of NOCS and SPD in comparison test of public data set CAMERA and REAL. It has obvious advantages especially under the situation that large hardware and data base is not feasible.

KEYWORDS: Transformer, pyramid pool, efficient self-attention mechanism, cross-modal fusion, 6-DoF pose estimation.

1. Introduction

The 6-DoF pose estimation of objects is widely used, including augmented reality, robot operation, and unmanned driving. In the field of augmented reality, virtual elements can be superimposed on the object by using the pose of the object, and the relative pose of the object remains unchanged with the movement of the object. In the field of robotics, with the maturity of simultaneous localization and mapping and other technologies, robots have been able to position well in space. Meanwhile, 6-DoF pose estimation technology is needed to locate objects to help robots interact with objects. In the field of autonomous driving, object pose estimation technology can sense other traffic participants and obstacles to provide information needed for decision-making. The result of pose estimation will affect the subsequent operation, and the low precision estimation result will lead to the failure of the later operation and planning task.

According to generalization, the existing methods can be divided into instance level and class level object 6-DoF pose estimation methods, which are also divided into three different algorithms based on correspondence, template and deep learning. The algorithm has the following problems, the corresponding algorithm relies on rich texture features or prominent shape features, so it is not suitable for objects with weak texture or not obvious shape features. To deal with the problems of illumination change, background clutter, and mutual occlusion between target objects, 2D detection frame information is generated by expanding SSD algorithm to infer object pose [15], or the 6-D pose of object is directly output by fusion channel spatial attention network CSA6D [11]. Such direct regression algorithm reduces the processing amount. However, the accuracy of the network estimation is not optimal in the application backgrounds such as inferior detection in factories and unmanned driving. Although the method based on deep learning to directly regression 6-D pose from a large number of data has high accuracy, it can only be used for instance-level object pose estimation with poor generalization [3, 10-11]. The algorithm based on template matching [8, 12] needs to increase the number of templates to improve the detection accuracy, but the cost is that the operation takes a long time and cannot realize real-time application. In short, the task of 6-DoF pose estimation is complicated, which needs to identify the class of the object, determine the area of the object in the image, and estimate the 6-DoF pose of the object. On the other hand, in order for the technology to be really used in practical tasks, it needs to deal with lighting changes, distance changes, background clutter, occlusion and hardware cost limitation. The existing methods are not sufficient for the situation that requires a high detection accuracy but only allows low hardware cost.

The existing approaches for 6-DoF pose estimation using RGB-D images can be categorized into three groups based on the manner in which the two sources of information are processed.

First kind, at the initial phase of feature extraction, the 6-DoF pose estimation is accomplished by merging the RGB and depth data. For example, the depth map is regarded as new channel information connected to the RGB channel and input to the CNN network for pose estimation. Second kind, RGB and depth information are used separately. Firstly stage, RGB images are used to predict rough 6D target pose, such as PoseCNN [7], BB8 [17], SSD-6D [15], YOLO-6D [1], PVNet (Peng et al., 2022), etc., and then ICP algorithm is used to refine target pose using depth information. Third kind, the fusion of color and depth information is designed to be postponed to feature extraction stage. Due to the difference over the data structures between the input color and depth information existed in distinct Spaces, the two heterogeneous information sources need to be processed separately, and geometric features and color features should be extracted on the premise of keeping the original data structure unchanged. In view of the above existing information source processing methods and the existing problems in the algorithm, we propose a novel attention-based multi-scale network. The core of the method is to use the attention mechanism to efficiently extract and fuse color information and geometric information, and use the multi-scale network structure to extract multi-scale dense features of different receptive fields containing target context information. Complete the 6-DoF pose estimation of object.

In this work, we proposed a low hardware cost 6-DoF network to improve the accuracy of object detection and recognition. In the cross-modal fusion procedure, we merge color and depth information at a basic level before utilizing the compact pyramid refinement (CPR) module to efficiently merge depth features across different levels, while using the implicit depth recovery technology (IDR) to enhance feature learning; The multi-head self-attention module of vision Transformer is encapsulated in the pyramid pool, which makes the scene understanding performance of backbone network P2T significantly superior to that of networks based on convolutional neural network and Transformer. Using P2T as backbone network, the compact pyramid refinement (CPR) module and implicit deep recovery (IDR) technique are used to optimize the 6-DoF pose estimation method for class-level objects, which improves the robustness and accuracy.

2. Related Work

2.1. Vision Transformer

Efficient modeling of multi-scale information is a key step in computer vision tasks. For example, semantic segmentation tasks require local details when determining object boundaries, whole-object level information when identifying object categories, and sometimes even a larger range of surrounding environment information to help make robust judgments. In addition to semantic segmentation, almost all visual tasks such as object detection, instance segmentation and object tracking require powerful multi-scale information modeling capabilities. In multi-scale modeling, because of the inherent complexity of large-scale models, it is very difficult to automatically learn the representation of large-scale features, and it is not very efficient to improve the receptive field range by stacking convolutional layers. The original purpose of Transformer is to solve machine translation tasks in the field of natural language processing [14], which is used to improve visual tasks due to its advantage of global modeling of dependencies between vocabularies through self-attention mechanism. Its core is to compute matrix multiplication of three sets of values of Query, Key and Value obtained by linear transformation of input sequence X to realize self-attention modeling of input sequence X. For example, DETR uses convolutional neural networks to extract two-dimensional features of images, flatten them and then feed them into Transformer [1], [22], [12], [7]. ViT [5] regards image blocks as lexical characters and works with Transformer network to directly process image classification tasks. PVT [18] and MViT [6] use single-layer pooling to reduce the number of words needed to compute a multi-head self-attention module, and none of the above algorithms seem powerful enough.

2.2. Pyramid-pool Transformer

Pooling is the one with the lowest computation cost in the basic operation of building multi-scale, and Pyramid Pooling should be the first step for high-efficiency multi-scale feature modeling. Pyramid pooling is applied to the vision Transformer backbone network, which not only reduces sequence length, but also learns more effective context features [20]. Therefore, compared with single-layer pooling, pyramid pooling is more efficient and can better calculate the self-attention relationship in multi-head self-attention module. The core of Pyramid Pooling Transformer (P2T) backbone network adopted in this paper [19] is to encapsulate pyramid pooling into multi-head self-attention module, which can improve the model's multi-scale expression ability while reducing the two matrix dimensions of K and V. The feature dimensions are compressed by average pooling of different scales, and the feature sequences after pooling are concatenated. Because multi-scale information is involved, the operation of positional embedding is carried out at the same time in order to avoid the quadratic change of the embedding in the multi-scale pooling. The location embedded here is implemented using DWConv. By controlling the proportion of pooling, the length of the spliced sequence after multi-scale pooling is slightly less than 1/64 of the original sequence length N, so that the calculation amount of Attention matrix can be well controlled, and the calculation process is more efficient. Pyramid pooling not only increases the computational efficiency, but also provides stronger multi-scale information processing capability. P2T is significantly superior to convolutional neural network and Transformer network in scene understanding tasks such as semantic segmentation, image classification, object detection, and instance segmentation.

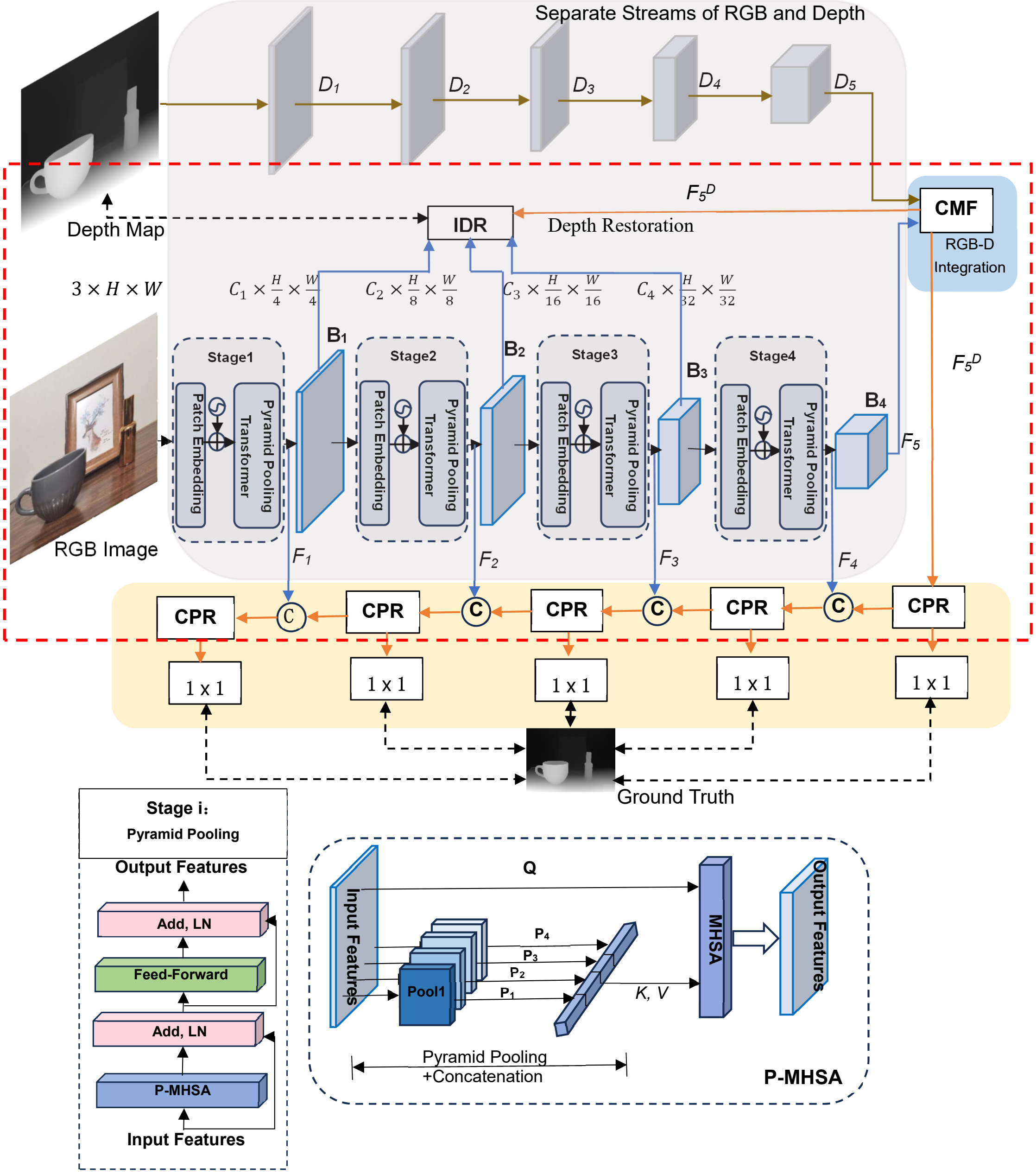
2.3. Multi-head Self-attention (MHSA) Module

Self-Attention mechanism, proposed by Bengio's team in 2014, has been widely used in various fields of deep learning in recent years, such as capturing receptive fields on images in the direction of computer vision, and locating key tokens or features in NLP.

The BERT algorithm proposed by the Google team has achieved amazing results in 11 tasks of NLP, The Transformer Network of BERT algorithm is completely composed of Attention mechanism, and its Transformer is composed of self-Attention and Feed Forward Neural Network only. However, the mecha-

Figure 1

Overall flow of P2T-Net network architecture and cross-modal fusion and Structure diagram of multi-head self-attention mechanism based on pyramid pooling(P-MHSA)



nism of multi-headed attention further improves the self-attention layer. P-MHSA, a multi-head self-attention module based on pooling, was adopted in this paper. Pyramid pooling was used to reduce the two matrix dimensions of K and V, while improving the multi-scale representation of the model. Its structure was shown in Figure 1.

2.4. Cross-mode Fusion and 6-DoF Pose Estimation

6-DoF pose estimation algorithm based on RGB data. At present, a large number of 6-DoF pose estimation networks based on RGB images have been studied. Since RGB images are usually represented in the form of three-dimensional arrays, they are suitable for Convolutional Neural networks (CNNS) to extract features. According to the different tasks performed by the network, the 6-DoF pose estimation network of the target object can be divided into direct regression and indirect regression. The 6-DoF pose estimation algorithm of indirect regression is mainly based on the commonly used neural network framework, which predicts the two-dimensional coordinates of key points in the input RGB image, obtains the correspondence between the two-dimensional coordinates of key points and the three-dimensional coordinates of key points in the target model, and then uses the PnP algorithm to solve the 6-DoF pose estimation of the target object. The 6-DoF pose estimation algorithm of direct regression adopts an end-to-end architecture to directly return the input RGB image to 6-DoF pose estimation, which is simpler and faster than the indirect regression method. Direct regression and indirect regression methods have better performance than traditional methods in terms of speed and robustness, but these methods learn the pose of the target object from the color features, and do not use spatial information, so the accuracy of pose estimation is still limited.

6-DoF pose estimation algorithm based on RGB-D data. The methods for estimating 6-DoF pose estimation based on RGB-D images are mainly divided into three categories according to the way in which the two information sources are processed. First, the RGB and depth information are fused at the early stage of feature extraction to complete the 6-DoF target pose estimation. For example, [11-12] the depth map is regarded as new channel information connected to the RGB channel and input to the CNN network for pose estimation. Second, RGB and depth information are used separately. In the first stage, RGB images are used to predict rough 6-DoF target pose estimation, such as PoseCNN, BB8, SSD-6D, YOLO-6D [6], PVNet (Peng et al., 2022), etc., and then ICP algorithm is used to refine target pose using depth information. The third method is to fuse color information and depth information in the later stage of feature extraction. Since the input RGB information and depth information exist in different Spaces and have different data structures, the two heterogeneous information sources need to be processed separately, and geometric features and color features should be extracted on the premise of retaining the original structure of the data.

3. Method

In this section, we first introduce the general idea of P2T-Net network structure in Section 3.1, and show the multi-head self-attention module based on pooling, which performs semantic segmentation of input RGB image and depth image to obtain the RGB information and depth information of the region of interest. In Section 3.2, we show cross-modal fusion of RGB features and depth features to extract multi-scale and multi-modal features for 6D pose estimation of target objects. Implicit depth recovery and compact pyramid thinning are described in Section 3.3. In Section 3.4 we cover loss functions.

3.1. Overview

Figure 1 indicates the P2T-Net network architecture. We extract features separately using RGB information flow and depth information flow respectively.

RGB information flow. For RGB information flow, similar to CNN's classic framework ResNet, the framework structure adopted in this paper includes four stages to generate feature maps of different sizes, each of which includes the patch Embedding layer and Pyramid Pooling Transformer layer. First, P2T-Net splits the input natural color image into $\frac{H}{4} \times \frac{W}{4}$ blocks, each flattened to 48 elements, and these flattened image blocks are input into an image block coding module containing a layer of linear projection with learnable position coding.

For the sake of simplicity, we discuss a set of self-attention modules without loss of generality. We used pyramid pooling to reduce the two matrix dimensions of K and V while improving the multi-scale representation of the model, as shown in Table 1.

Table 1

Pyramid Pooling based on MHSA

Traditional MHSA	Pyramid Pooling based MHSA
$(Q,K,V)=(XW^q,XW^k,XW^v)$	$(Q,K,V)=(XW^q,PW^k,PW^v)$
Complexity:$O(NC^2+N^2C)$ N: Sequence length, C: feature dimension	$Attention(Q,K,V)=$ $Softmax(\frac{Q\times K^T}{\sqrt{d_K}})\times V$ Computational complexity: $O((N+2M)C^2+2NMC)$ M: length of pooled feature $M\approx N/8^2$

In P MHSA, X as the input is adjusted and rectified into a two-dimensional space, and we apply multiple average pool layers of different proportions to X to generate a pyramid feature map. The details are as follows:

Pooling:

$$P_1 = AvgPool_1(X)$$
$$P_2 = AvgPool_2(X)$$
$$P_n = AvgPool_n(X)$$

Concatenate:

$$P_i^{enc} = DWConv(P_i)+P_i, i=1,2,...,n$$
$$P = LNorm(Concat(P_i^{enc}...)).$$

Here there is a convolution layer with step size 2 after each stage, so after each stage, the resolution of the feature map will be downsampled to half of the original. We represent the output mapping of the five stages as: F_1, F_2, F_3, F_4, F_5, with steps of 2, 22, 23, 24,25, respectively.

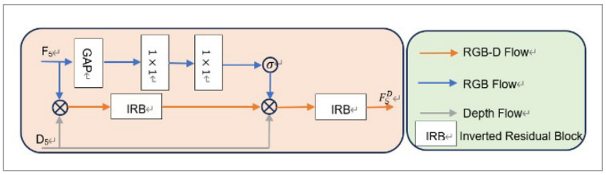
Deep information flow. Similar to the RGB flow, the deep flow has five phases with the same step size. For the reason that, compared with the corresponding color image, the depth map image reduces the semantic information, its convolutional blocks are used less than the RGB information flow. Depth maps distinguish foreground objects from backgrounds by interpreting the spatial cues in RGB images, which works more effectively when applying to images containing complex textures and scenes. In the two inverted residual blocks used in the previous stage, we first extend the feature graph M times along the channel dimension by 1 × 1 convolution into a depth-separable 3 × 3 convolution with the number of input channels and output channels remaining the same. The feature channel is then compressed to the original $\frac{1}{M}$ by another 1 × 1 convolution, and each convolution has a BN layer and a ReLU layer, except the last one that passes through a 1 × 1 convolution. The output characteristic graphs of the depth information flow in the five stages are D_1, D_2, D_3, D_4, D_5, in which the first four channels have 16, 32, 64 and 96 channels respectively. During the process, the quantity and step length of D_5 as well as F_5 channels are the same.

3.2. Cross-modal Fusion of Color Information and Depth Information

To reduce the computing cost caused by a large feature resolution, we integrated RGB features and depth features at the coarser level by applying a cross-modal fusion (CMF) module to. Depth maps convey a smooth prior of deep regions that are able to roughly represent the shape and structure of a complete target or object. Therefore, multiplying the depth feature with the semantic feature of RGB information to enhance the semantic feature of RGB information is a strong regularization operation. Other modes of interaction can only be based on the equal treatment of the characteristics of the two characteristics. These operations are just orthogonal to our goal and not consistent with it [21].

Figure 2

Cross-modal Fusion module (CMF) structure



We start by fusing the extracted RGB feature F_5 and depth feature D_5, which utilize the output of RGB information flow and depth information flow, in order to generate the RGB-D feature F_5^D. This process can be expressed as:

$$\tau = IRB(F_5 \otimes D_5)$$
$$v = \sigma(FF_2(R_E LU(FF_1(GAP(F_5)))))$$
$$F_5^D = IRB(v \otimes \tau \otimes D_5).$$

3.3. Depth Recovery (IDR) and Pyramid Thinning (CPR)

In the IDR module shown in Figure 3(a), a 1 × 1 convolution is applied to reduce the channel count for F_1, F_2, F_3, F_4, F_5^D to 256. Afterwards, adjust the size of the output feature maps to be consistent with F_4 and connect them. Four consecutive IRB modules are then used to fuse the multilevel features to generate significant multi-scale features. Finally, we utilize a basic 1 × 1 convolution to convert the fused feature map into a map with only one channel. By employing a standard sigmoid function and bilinear upsampling, the restoration process yields a depth map that matches the dimensions of the original input image.

As shown in Figure 3(b) of the CPR module, assuming that the input of the CPR module is χ, CPR uses 1 × 1 convolution to increase the input channel number with a multiplication of M. Three 3 × 3 depth separable convolution with expansion rates of 1, 2 and 3 were paralleled to achieve the fusion over several scales. This process is able to be expressed as:

$$\chi_1 = Conv_{1\times1}(\chi)$$
$$\chi_2^{d_1} = Conv_{3\times3}^{d_1}(\chi_1)$$
$$\chi_2^{d_2} = Conv_{3\times3}^{d_2}(\chi_1)$$
$$\chi_2^{d_3} = Conv_{3\times3}^{d_3}(\chi_1)$$
$$\chi_2 = \mathrm{Re}\,LU(BN(\chi_2^{d_1} + \chi_2^{d_2} + \chi_2^{d_3})),$$

where d_1, d_2, d_3 are the expansion rates, which are 1, 2, 3 respectively.

$$\chi_3 = Conv_{1\times1}(\chi_2) + \chi$$
$$y = v' \otimes Conv_{1\times1}(\chi_3)$$

The adjusted fused features are determined by global context information according to formula above.

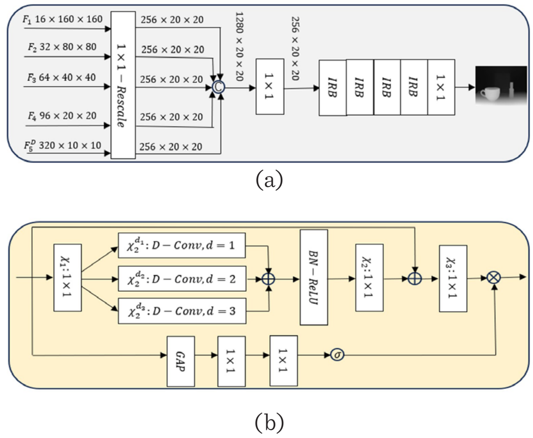
In each decoder stage, initially halve the number of channels in the two feature maps from the top decoder and corresponding encoder stage using 1 × 1 convolution. Then concatenate these results by channel dimension, followed by feature fusion using the CPR module.

3.4. Loss Function

In order to solve the problems of occlusion and truncation, this paper conducts intensive regression on the extracted multi-scale intensive discriminant features and obtains the intensive prediction results. We defined the loss estimated by pose as the distance between the sampling points on the target model after the real pose transformation and the predicted pose transformation with multi-scale feature intensive regression, and the loss function of each intensive prediction result is as follows:

Figure 3

Implicit Depth Restoration (IDR) (a) and Compact Pyramid Refinement (CPR) (b)



$$L_i^p = \frac{1}{M}\sum_j \| (Rx_j + t) - (\tilde{R}_i\, x_j + \tilde{t}_i) \|,$$

where, x_j is the point of the target 3D model, p=$[R|t]$ is the marked real pose, $\tilde{p}_i = [\tilde{R}_i|\tilde{t}_i]$ is the pose predicted from the network dense feature, and M is the number of target 3D model points. However, this specific loss function is unable to effectively deal with objects that possess symmetry, because there are unpredicted numbers of correct rotations for a symmetric object. To avoid penalizing our network by reverting to one of the optional correct rotations, we use a loss function for symmetric objects, which defines the loss as the distance between the sampling point on the target model after the true pose transformation and the nearest point after the predicted pose transformation. In this case, the densely predicted loss function is:

$$L_i^p = \frac{1}{M}\sum_j \min_{0<k<M} \| (Rx_j + t) - (\tilde{R}_i\, x_k + \tilde{t}_i) \|.$$

For intensive prediction results, the best result needs to be selected as the network output, so our network also needs to learn to independently select the most likely correct pose $[\tilde{R}_i|\tilde{t}_i]$. For this reason, we add the self-supervised dense confidence c_i to the loss function of intensive prediction for weighting, and add a confidence regularization term. Make the pose corresponding to the highest confidence be the pose of the final network output:

$$L_s = \frac{1}{N}\sum_i (L_i^p c_i - \omega \log(c_i)),$$

where N is the pose number of dense predictions, and ω is a confidence regularization term of a balanced hyperparameter that provides a high penalty for losses with low confidence. L_s is the loss function corresponding to multimodal features of different scales. Thus, the loss function for minimizing this loss to estimate the pose is:

$$L = \sum_s L_s,$$

where S is the number of multimodal features of different sizes, and S set to 3 represents the three scales of multimodal features.

4. Experimental Results and Analysis

In the experiment platform, the hardware includes: Intel Xeon Silver 4210R, 24G RAM, and RTX3090 GPU; The software includes: 64-bit Ubuntu operating system, version 18.04LTS, Python3.6, Cuda11.0, Cudnn8.1.1, Anaconda 3, and Pytorch1.7.1.

4.1. Data Set

In this paper, Linemod, CAMERA, REAL and YCB-Video, four authoritative public data sets in the field of 6-DoF pose estimation of class-level objects, are used for training and testing. In this paper, the performance of the pose estimation network is verified on these four data sets and compared with the existing algorithms.

The Linemod dataset includes 13 kinds of low-texture objects with complex background, occlusion and truncation, and noise in the depth data. In order to properly evaluate the comparison with the previous work, we used 85% of the images in the dataset as training images, while 15% were used for testing.

CAMERA is a composite dataset containing 300K composite images; the training set is 275K, including 1085 different object instances; The remaining 25K is used as a test set, containing 184 different object instances.

REAL is a supplement to CAMERA. 4,300 images in 7 scenarios were used for training; 2,750 images in 6 scenarios were used for testing; The training and test sets contain three different object instances per class.

YCB-Video dataset contains 92 videos. Each video randomly selects the target from 21 objects for arbitrary placement, then moves the camera to finish shooting, and finally selects the video frame as the data set. The whole dataset contains 133827 frames of images, each including RGB information and depth information, and 6-D pose annotation is carried out in a semi-automatic way. In order to compare with the existing work, According to our previous work, 16,989 frames of images were randomly selected from 80 videos together with 80,000 frames of composite images as the training set, and then 2949 frames were randomly selected from the remaining 12 videos as the training set. Our experimental results are compared with existing advanced pose estimation networks using the semantic segmentation results provided in the YCB-Video dataset. The YCB-Video dataset has objects and occlusion conditions of various shape and texture levels under different lighting, making it ideal for evaluating the robustness of methods.

4.2. Evaluation Indicators

In this paper, the 6-DoF pose estimation and mean distance (ADD) measurement are utilized as a standard to rate the effectiveness of 6DoF pose estimation for class-level objects.

n° m cm (average precision (mAP)) means that when the rotation error of the estimated pose is less than n° and the translation error is less than m cm, the pose is considered correct. It is worth noting that for symmetrical objects (such as bottles, bowls, and jars), any prediction of rotation along the axis of symmetry is considered correct. In addition, when the handle of the mug is not visible, it is considered to be a symmetric object, and the reverse is an asymmetric object. In this paper, the mAP of intersection over union (3D IoU), whose accuracy threshold is 75%, is abbreviated as 3D75 [11].

Average point distance Measurement method (ADD). ADD refers to the average Euclidean distance between the points on the CAD model of the target object after real rotation $\tilde{R}$ and translation $\tilde{t}$ and corresponding to the predicted rotation R and translation t:

$$ADD=\frac{1}{m}\sum_{x\in M}\|(Rx+t)-(\tilde{R}x+\tilde{t})\|,$$

where m is the number of model points and M is the model point set. In Linemod dataset, we set the evaluation threshold of ADD as 10% of the diameter of the target object, that is, when the ADD score of the network output pose is less than 10% of the diameter of the target object, we call this pose the correct pose.

4.3. Experimental Conclusion

On Linemod dataset, ADD and ADD-S evaluation indexes are used for quantitative evaluation. If ADD(-S) is less than 10% of the object diameter, the attitude estimation is considered to be correct, and the ADD(-S) score is defined as the correct ratio of attitude estimation. Where ADD(-S) represents the ADD of an asymmetric object and the Add-S metric of a symmetric object, respectively. In the experiment, we compare the proposed network with the results of existing advanced algorithms BB8, PVNet, PoseCNN, SSD-6D and DenseFusion. Table 2 shows the evaluation results of each algorithm on all 13 objects in Linemod dataset, in which the average ADD(-S) score of our network among the 13 objects is 94.6, showing advanced performance.

In the test on CAMERA25, the value of 3D$_{75}$ mAP in this paper is 86.3%, which is higher than the baseline method NOCS (He et al., 2019), and SPD [13] is 16.8% and 3.2%, respectively, which is higher than SAR-NET [11] 7.3%. In this method, mAP of 10°2cm and 10°5cm reached 81.4% and 87.1%, which exceeded SAR-Net [11] by 6.1% and 6.8%. At the more stringent standards of 5°2cm and 5°5cm, the proposed method still has a significant advantage of 69.2% and 72.9%, which exceeds that of SAR-Net [11]2.5% and 2.0%.

In the test on REAL275, the 3D$_{75}$ mAP of this paper is 65.1%, which exceeds the baseline method NOCS (He et al., 2019) and SPD [13] by 35% and 11.9%, respectively. In this paper, mAP of 10°2cm and 110°5cm reaches 61.4% and 71.2%, which exceeds SAR-Net [11] by 11.1% and 2.9%. The superiority of the proposed method is also demonstrated in the more stringent standards of 5°2cm and 5°5cm, which are 7.6% and 1.0% higher than SAR-Net [11] and 3.0% and 3.4% higher than SGPA [2]. The results above show that the proposed method has strong universality and can accurately estimate the pose of new objects of the

Table 2
Quantitative evaluation of ADD score on Linemod dataset

Method	BB8	PVNet	PoseCNN +DeepIM	SSD-6D +ICP	Dense-Fusion	Dense-Fusion +refinement	ours
Data	RGB			RGB-D			
ape	40.4	43.6	77	65	79.5	**92.3**	90.1
bench	91.8	**99.9**	97.5	80	84.2	93.2	94.5
camera	55.7	86.9	93.5	78	76.5	94.4	**95.9**
can	64.1	95.5	**96.5**	86	86.6	93.1	95.3
cat	62.6	79.3	82.1	70	88.8	**96.5**	93.7
driller	74.4	**96.4**	95	73	77.7	87	91.3
duck	44.3	52.6	77.7	66	76.3	**92.3**	89.4
eggbox	57.8	99.2	97.1	**100**	99.9	99.8	**100**
glue	41.2	95.7	99.4	**100**	99.4	**100**	**100**
hole	67.2	81.9	52.8	49	79	**92.1**	**92.6**
iron	84.7	**98.9**	98.3	78	92.1	97	96.5
lamp	76.5	**99.3**	97.5	73	92.3	95.3	95.2
phone	54	92.4	87.7	79	88	92.8	94.8
MEAN	62.7	86.3	88.6	76.7	86.2	94.3	**94.6**

Note: The bold part is the optimal value of the line.

Table 3

Comparative results on REAL275 and CAMERA25 datasets

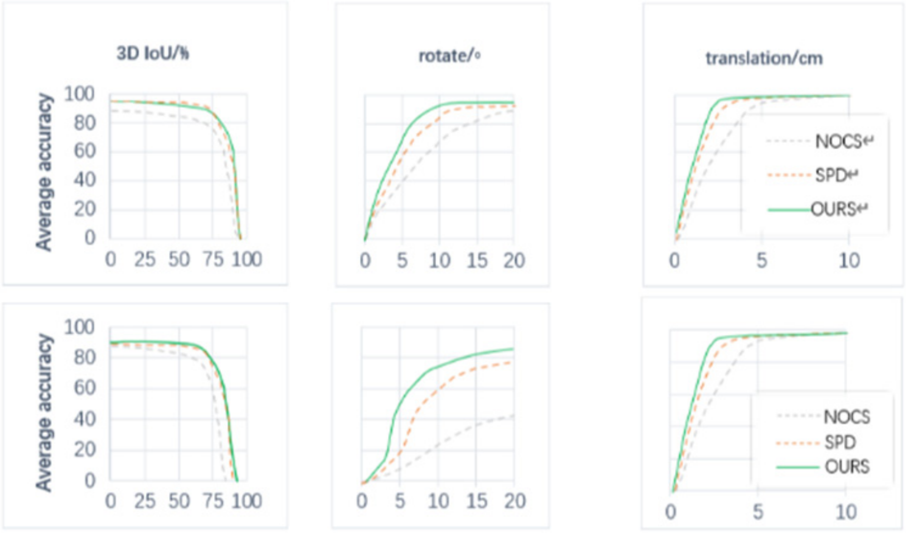
Data set	Method	mAP				
		$3D_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
REAL275	NOCS (He et al., 2019)	30.1	7.2	10.0	13.8	25.2
	SPD[13]	53.2	19.3	21.4	43.2	54.1
	DualPoseNet[10]	62.2	29.3	35.9	50.0	66.8
	SGPA[2]	64.8	36.2	39.9	61.5	70.7
	CenterSnap[8]	-	-	29.1	-	64.3
	SAR-Net[11]	62.4	31.6	42.3	50.3	68.3
	iCaps[4]	-	-	22.3	-	-
	FS-Net[3]	63.5	-	28.2	-	60.8
	CR-Net[17]	55.9	27.8	34.3	47.2	60.8
	Ours	65.1	39.2	43.3	61.4	71.2
CAMERA25	NOCS (He et al., 2019)	69.5	32.3	40.9	48.2	64.6
	SPD[13]	83.1	54.3	59.0	73.3	81.5
	DualPoseNet[10]	86.4	64.7	70.7	77.2	84.7
	SGPA[2]	85.6	68.3	72.3	81.2	86.8
	CenterSnap[8]	-	-	66.2	-	81.3
	SAR-Net[11]	79.0	66.7	70.9	75.3	80.3
	Ours	86.3	69.2	72.9	81.4	87.1

same class, which is in line with the needs of practical applications. The quantitative comparison with cutting-edge methods shows that the proposed method has advanced performance. Table 3 compares the performance of each method on the data set CAMERA25 and REAL275.

Figure 4 clearly shows the performance comparison curve between the proposed method and the two baseline methods, NOCS and SPD. It can be seen from the figure that the proposed method has a significant advantage when the 3D IOU threshold is greater than 75%, and the estimation of rotation and migration is

Figure 4

Comparison of the performance curves of CAMERA25 (top) and REAL275 (bottom) for different methods



significantly improved compared with the baseline method.

In Table 4, we compare the speed of our approach to the most advanced methods based on RGB and RGB-D data. We can see that our method performs 6% better in ADD(-S) accuracy than PoseCNN+DeepIM and runs almost 4 times faster. Compared to DenseFusion+refinement, our method achieves 0.3% higher ADD(-S) score and is 25% faster.

Table 4

The speed (frames per second, FPS) of different methods for pose estimation

Method	PVNeT	PoseCNN +DeepIM	Densefusion +refinement	OURS
Speed(FPS)	25	5	16	20

Table 5 shows the evaluation results of 21 objects in the YCB-Video dataset using semantic segmentation of PoseCNN. If ADD(-S) is less than the mean distance threshold, the pose estimate is considered correct, and the ADD(-S) score is defined as the percentage of pose estimates that are correct. Two evaluation indicators are used here to measure the effectiveness of our approach. One is the area under the ADD(-S) score-threshold curve (AUC), whose threshold changes from 0cm to 10cm. Another indicator is the ADD(-S) score, which has a threshold of 2 cm.

In Table 5, we can see that our method outperforms Densefusion and PoseCNN+ICP by 2.2% and 3.8% on the second metric, and by 1.2% on the first metric. Since the YCB-Video dataset has objects under different lighting and occlusion conditions, our method is

Table 5

Quantitative evaluation of ADD(-S) AUC and ADD(-S) scores (<2cm) on the YCB-Video Dataset

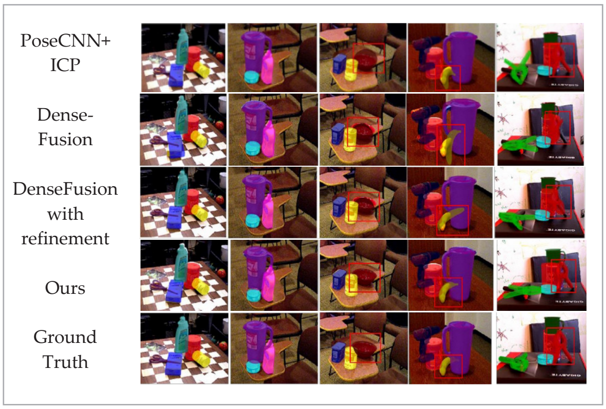
Method	PoseCNN+ICP		DenseFusion		OURS	
	AUC	<2cm	AUC	<2cm	AUC	<2cm
master_chef_can	68.1	51.1	70.7	70.7	67.7	65.6
cracker_box	83.4	73.3	86.8	88.6	89.9	95.2
sugar_box	97.5	99.5	90.8	96.8	97.1	99.5
tomato_soup_can	81.8	76.6	84.7	82.8	85.2	84.4
mustard_bottle	98.0	98.6	90.9	94.1	90.7	93.8
tuna_fish_can	83.9	72.1	79.5	58.5	79.5	63.0
pudding_box	96.6	100.0	89.4	94.4	89.3	93.1
gelatin_box	98.1	100.0	95.7	100.0	93.5	100.0
potted_meat_can	83.5	77.9	79.6	76.9	81.3	77.3
banana	91.9	88.1	76.8	60.2	80.0	64.5
pitcher_base	96.9	97.7	87.1	87.2	91.0	91.4
bleach_cleaner	92.5	92.7	87.5	85.4	88.3	89.1
bowl	81.0	54.9	86.1	61.3	93.1	98.9
mug	81.1	55.2	83.9	80.5	83.3	73.0
power_drill	97.7	99.2	83.7	83.1	82.6	77.0
wood_block	87.6	80.2	89.4	98.8	91.0	99.1
scissor	78.4	49.2	77.1	50.8	77.0	49.2
large_marker	85.3	87.2	89.1	90.6	91.1	98.0
large_clamp	75.2	74.9	71.5	78.0	71.5	77.3
extra_large_clamp	64.4	48.8	70.1	72.0	68.3	68.7
foam_brick	97.2	100	92.2	100.0	95.1	100.0
MEAN	86.6	79.9	83.9	81.5	85.1	83.8

more robust to lighting and occlusion changes. Objects in the YCB-Video dataset can be classified and analyzed based on their texture and geometric features. In “cracker_box”, “sugar_box”, “tomato_soup_can”, and “wood_block”. On objects with smooth surfaces and rich textures, our method has better performance than SOTA method. The ADD(-S) performance is greatly improved on smooth and untextured objects such as “bowl” and “large_marker”, indicating that the type of context information extraction target of these objects can improve the accuracy of pose estimation.

Figure 5 visualises PoseCNN+ICP, DenseFusion, DenseFusion + refinement, our network pose prediction results and marked real poses, and transforms objects of different colors into predicted poses. The red boxes highlight objects where our method performs significantly better than the other three methods, from left to right “bowl,” “banana,” and “clamp.” It can be seen that all networks are occlusion-resistant on smooth, untextured objects, such as “tomato_soup_can”, “mustard_bottle” in class 21 objects in the YCB-Video dataset. For “mug”, “banana” and “pudding_box”, the attitude estimated by our method is closest to the real marked pose. This is because the object context considered in our proposed method enriches the information of untextured objects and helps to improve the accuracy of pose estimation. For the challenge of poor segmentation results, our network can position these objects (such as “clamps” and “scissors”) well through prediction, and can further improve the rotation transformation performance of these objects.

Figure 5

Some qualitative results on the YCB-Video dataset



In order to further demonstrate the performance of this method in dealing with similar new objects,

Table 6

RGB - D fusion and melting IDR branch research

No.	Features for fusion					IDR	F_β^{max}	MAE
	C_1	C_2	C_3	C_4	C_5			
1	√	√	√	√	√	√	0.899	0.052
2	√	√				√	0.894	0.050
3	√	√	√			√	0.897	0.047
4		√	√	√		√	0.902	0.048
5			√	√	√	√	0.902	0.046
6					√	√	0.906	0.045
7	√	√	√	√	√		0.895	0.047
8	√	√					0.892	0.049
9	√	√	√				0.896	0.048
10		√	√	√			0.895	0.048
11			√	√	√		0.898	0.048
12					√		0.896	0.047
13							0.887	0.052

the method is tested against 4 types of interference scenes, 5 types of objects, and 6 object examples. The four types of interference include illumination change, distance change, background clutter and occlusion. The 5 categories of objects include cameras, bottles, mugs, bowls, and cans. Figure 6 shows some of the test results in real world scenarios, showing robust and excellent performance.

Table 6 shows the results of RGB-D fusion at different stages. When IDR branch training is used, the model fuses RGB features and depth features at the coarsest level, and No.6 performs best. In the case of No IDR branch, the model fuses the RGB and depth features of the three layers, and the training results of No.11 are the best. In most cases, IDR branches significantly improve performance, which validates the effectiveness of our proposed combination of fusion strategy and IDR branches. Although the earlier fusion strategy was more efficient in linking the input RGB image with the depth map at the input stage, Table 7 shows that our initial fusion strategy was significantly superior to it. To ensure accuracy and efficiency, we blend RGB and snake venom features at the coarsest level. In ablation studies, we mainly used F_β^{max} and MAE as indicators.

Table 7

Comparison of fusion strategies

Metric	Single Stream		Two Streams	
	IDR√	IDRx	IDR√	IDRx
F_β^{max}	0.900	0.894	0.906	0.896
MAE	0.048	0.051	0.045	0.047

Table 8 shows the trial results of CPR, in which different inflation strategies were used. We tested the default setting (No. 1), a single convolution with different expansion rates (No. 2-4), and multiple convolution with sparse combinations of expansion rates (No. 5-6), respectively. The compression spread rate (1, 2, 3) of the default setting is significantly better than the other Settings, illustrating the effectiveness of CPR.

5. Conclusion

In order to achieve accurate and fast 6-DoF pose estimation of objects and solve the problem of pose estimation under interference scenes, in this paper, we creatively proposed a P2T-Net-network based method combining both RGB channel and depth channel in analysis. In the process of cross-modal fusion, RGB information and depth information are first fused at the rough level, and then the compact pyramid refinement (CPR) module is used to effectively integrate multi-level depth features. At the same time, the implicit deep recovery technology (IDR) is used to strengthen feature learning, effectively aggregate details and overall information, highlight important information, and improve accuracy and speed. At the same time, through experiments on public data sets Linemod, CAMERA and REAL, the mean average accuracy of the proposed method is better than NOCS, SPD, SGPA and some other mainstream 6-DoF pose estimation methods for class-level objects. Further experiments show that the proposed method can accurately estimate the 6-DoF pose of objects in the scene with illumination variation, distance variation, background clutter, occlusion and other disturbances. Using P2T as backbone network, the compact pyramid refinement (CPR) module and implicit deep recovery (IDR) technique are used to optimize the 6DoF pose estimation method for class-level objects, which improves the robustness and accuracy. In the future, we will further improve our method to apply to pose estimation of transparent objects, which are more common in family scenes and have practical value.

Table 8

CPR expansion rate of melting

No.	1	2	3	4	5	6
Expansion rate	1,2,3	1	2	3	1,3,6	1,4,8
F_β^{max}	0.906	0.901	0.892	0.897	0.903	0.900
MAE	0.045	0.047	0.049	0.047	0.046	0.048

Data sharing agreement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, author-ship, and/or publication of this article.

Funding

The author(s) received no financial support for the research.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. M., Zagoruyko, S. End-to-End Object Detection with Transformers, 2020. ArXiv (Cornell University). https://doi.org/10.1007/978-3-030-58452-8_13
2. Chen, K., Dou, Q. SGPA: Structure-Guided Prior Adaptation for Category-Level 6D Object Pose Estimation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. https://doi.org/10.1109/ICCV48922.2021.00277
3. Chen, W., Jia, X., Hyung Jin Chang, Duan, J., Shen, L., Leonardis, A. FS-Net: Fast Shape-based Network for Category-Level 6D Object Pose Estimation with Decoupled Rotation Mechanism. University of Birmingham Research Portal (University of Birmingham), 2021. https://doi.org/10.1109/CVPR46437.2021.00163
4. Deng, X., Geng, J., Bretl, T., Xiang, Y., Fox, D. iCaps: Iterative Category-Level Object Pose and Shape Estimation. IEEE Robotics and Automation Letters, 2022, 7(2), 1784-1791. https://doi.org/10.1109/LRA.2022.3142441
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021. ArXiv.org. https://doi.org/10.48550/arXiv.2010.11929
6. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C. Multiscale Vision Transformers. 2021. ArXiv.org. https://doi.org/10.1109/ICCV48922.2021.00675
7. Hu, J., Cao, L., Lu, Y., Zhang, S., Wang, Y., Li, K., Huang, F., Shao, L., Ji, R. ISTR: End-to-End Instance Segmentation with Transformers. ArXiv (Cornell University), 2021. https://doi.org/10.48550/arxiv.2105.00637
8. Irshad, M. Z., Kollar, T., Laskey, M., Stone, K., Kira, Z. CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and Categorical 6D Pose and Size Estimation. 2022 International Conference on Robotics and Automation (ICRA), 2022. https://doi.org/10.1109/ICRA46639.2022.9811799
9. Lin, H., Liu, Z., Cheang, C., Fu, Y., Guo, G., Xue, X. SAR-Net: Shape Alignment and Recovery Network for Category-level 6D Object Pose and Size Estimation. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. https://doi.org/10.1109/CVPR52688.2022.00659
10. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y. DualPoseNet: Category-level 6D Object Pose and Size Estimation Using Dual Pose Network with Refined Learning of Pose Consistency. ArXiv (Cornell University), 2021. https://doi.org/10.1109/ICCV48922.2021.00354
11. Liu, C., Sun, W., Zhang, K., Liu, J., Zhang, X., Fan, S. (Prior Geometry Guided Direct Regression Network for Monocular 6D Object Pose Estimation. 2022 41st Chinese Control Conference (CCC), 2022. https://doi.org/10.23919/CCC55666.2022.9901912
12. Liu, J., Sun, W., Liu, C., Zhang, X., Fan, S., Wu, W. HFF6D: Hierarchical Feature Fusion Network for Robust 6D Object Pose Tracking. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11), 7719-7731. https://doi.org/10.1109/TCSVT.2022.3181597
13. Tian, M., Ang, M. H., Lee, G. H. Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation. ArXiv (Cornell University), 2022. https://doi.org/10.48550/arxiv.2007.08454
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. Attention Is All You Need. 2017. ArXiv.org. https://doi.org/10.48550/arXiv.1706.03762
15. Wang, G., Manhardt, F., Tombari, F., Ji, X. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. https://doi.org/10.1109/CVPR46437.2021.01634

16. Wang, H., Zhu, Y., Adam, H., Yuille, A. L., Chen, L.-C. MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers. Computer Vision and Pattern Recognition, 2021. https://doi.org/10.1109/CVPR46437.2021.00542

17. Wang, J., Chen, K., Dou, Q. Category-Level 6D Object Pose Estimation via Cascaded Relation and Recurrent Reconstruction Networks. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021. https://doi.org/10.1109/IROS51168.2021.9636212

18. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. 2021. https://doi.org/10.1109/ICCV48922.2021.00061

19. Wu, Y.-H., Liu, Y., Zhan, X., Cheng, M. P2T: Pyramid Pooling Transformer for Scene Understanding. ArXiv (Cornell University), 2021. https://doi.org/10.48550/arxiv.2106.12011

20. Zheng, Q., Tian, X., Yu, Z., Jiang, N., Elhanashi, A., Saponara, S., Yu, R. Application of wavelet-packet transform driven deep learning method in PM2.5 concentration prediction: A case study of Qingdao, China. Sustainable Cities and Society, 2023, 92, 104486. https://doi.org/10.1016/j.scs.2023.104486

21. Zheng, Q., Tian, X., Yu, Z., Wang, H., Elhanashi, A., Saponara, S. DL-PR: Generalized automatic modulation classification method based on deep learning with priori regularization. Engineering Applications of Artificial Intelligence, 2023, 122, 106082. https://doi.org/10.1016/j.engappai.2023.106082

22. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. 2021, ArXiv.org. https://doi.org/10.48550/arXiv.2010.04159